DOCUMENT RESUME

ED 462 407 TM 033 669

AUTHOR Lei, Pui-Wa; Bassiri, Dina; Schultz, E. Matthew
TITLE Alternatives to the Grade Point Average as a Measure of Academic Achievement in College. ACT Research Report Series.
INSTITUTION American Coll. Testing Program, Iowa City, IA.
REPORT NO ACT-RR-2001-4
PUB DATE 2001-12-00
NOTE 39p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243-0168. Tel: 319-337-1110; Web site: http://www.act.org/research/reports/index.html.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *College Students; *Grade Point Average; Higher Education; Item Response Theory; Measurement Techniques; Rating Scales; *Student Evaluation
IDENTIFIERS Graded Response Model

ABSTRACT

College grade point average (GPA), a linear combination of assigned grades from different courses, is widely known to be an imperfect measure of student achievement. This unreliable measure decreases the predictive validity of college admission tests. Research has shown that adjusting course grades for differential grading practices improves predictive validity. Relative rankings of students on adjusted college GPAs are also more consistent with their course grade standings. These findings were replicated with course grade data from consecutive cohorts of two universities using four polytomous item response theory (IRT) and three linear models. In study 1, data were available for 1,255 students in the 1995 cohort and 1,796 students in the 1996 cohort. In the second study, data were available for 1,823 students for 1995 and 1,879 students for 1996. As had been done in previous studies, course parameter estimates and regression weights were cross-validated. Both same-sample and cross-validated alternative measures showed improvement over simple GPA. The rating scale and partial credit IRT models excelled on multiple correlations with admission test scores. The graded response IRT model was the most unstable across cohorts. Implications of these findings and limitations of the studies are discussed. An appendix shows formulas and number of course parameters for the models selected. (Contains 4 figures, 6 tables, and 24 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ACT Research Report Series 2001-4

ED 462 407

Alternatives to the Grade Point Average as a Measure of Academic Achievement in College

Pui-Wa Lei

Dina Bassiri

E. Matthew Schultz

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033669

BEST COPY AVAILABLE

ACT

December 2001

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

© 2001 by ACT, Inc. All rights reserved.

Alternatives to the Grade Point Average as a Measure of Academic Achievement in College

Pui-Wa Lei
Dina Bassiri
E. Matthew Schulz

Abstract

College GPA, a linear combination of assigned grades from different courses, is widely known to be an imperfect measure of student achievement. This unreliable measure decreases the predictive validity of college admission tests. Research has shown that adjusting course grades for differential grading practices improves predictive validity. Relative rankings of students on adjusted college GPAs are also more consistent with their course grade standings. These findings were replicated with course grade data from consecutive cohorts of two universities using 4 polytomous IRT and 3 linear models. Unlike previous studies, course parameter estimates and regression weights were cross-validated. Both same-sample and cross-validated alternative measures showed improvement over simple GPA. The rating scale and partial credit IRT models excelled on multiple correlations with admission test scores. The graded response IRT model was the most unstable across cohorts. Implications of these findings and limitations of the studies are discussed.

Alternatives to the Grade Point Average as a Measure of Academic Achievement in College

GPA, a linear combination of grades assigned in different courses, is not an ideal measure of student achievement because it reflects not only academic achievement, but also course taking strategies and instructor grading practices. Unless course selection is not allowed and all instructors are willing to adhere to a universal grading standard, GPAs for different students do not necessarily have the same meaning.

Several problems are associated with the use of GPA as a measure of academic achievement in college. One is that it is difficult to select candidates from different departments or different institutions for scholarship or employment purposes (e.g., Caulkins, et al., 1996). Different teachers/professors have different grading criteria according to their own perception of student achievement (Hoover, Roller, Liddell, Moore, McCarthy, and Hlebowitsh, 1999). Perhaps due to the composition of "like-minded" individuals, departments vary in their grading tendencies (Hoover, et al., 1999; Johnson, 1997). Grade point averages are, therefore, not strictly comparable among students, particularly across departments or majors.

The use of GPA as a measure of academic achievement also drives grade inflation. It has been suggested that instructors lower their standards in order to improve their course ratings by students. Students may shop for courses taught by leniently grading instructors (e.g., Johnson, 1997) or switch to departments that tend to give high grades (e.g., Young, 1993). As a result, grades may be raised without reflecting increased students' abilities, a phenomenon known as grade inflation (Bejar & Blew, 1981). Note that grade inflation is not necessary specific to the college level. Ziomek and Svec (1995) documented a similar phenomenon at the high school level.

Relying on GPA as a measure of academic achievement also makes it more difficult to evaluate college admissions tests (e.g., Young, 1993; Stricker, Rock, Burton, Muraki, & Jirele, 1994). Admissions tests are rightly expected to predict grades but should not be expected to predict whether a student will choose an easier curriculum.

To make course grades more comparable, a viable alternative to imposing a common grading standard on all instructors (or to denying course selection by students) is to adjust GPA for differential course difficulty (Caulkins, Larkey, and Wei, 1996). Adjusted-GPA, like GPA, represents student achievement on a single (unidimensional) scale and is constructed entirely from course grade data. An adjusted-GPA does not resolve the underlying problem of representing an inherent multidimensional domain with respect to the specific subject areas with a simple unidimensional measure. Other dimensions may also include non-cognitive characteristics such as attending class and turning in homework. However, it is better than GPA because it reduces the error arising from differential course-taking patterns and variation in course difficulty. It is hoped that by leveling course difficulty, incentives other than learning the course content will be discouraged in the long run. Immediate effects of adjusting GPA include improved predictive validity of college admission tests (Young, 1990a; Caulkins, et al., 1996; Johnson, 1997) and reduced differential predictive validity for gender (Young, 1991).

Most GPA adjustment methods operate on the premise that an achievement index should reflect relative course ranks of students who took the same classes (e.g., Caulkins, et al, 1996; Johnson, 1997), a unidimensional notion of achievement. These methods often adjust for the different grading stringency or difficulty level of the classes. Methods of particular interest to this study are summarized in the following section. For recent comprehensive reviews of

adjustment methods, see Young (1993) and Stricker, Rock, Burton, Muraki, and Jirele (1994), or a more distant one, see Linn (1966).

Grade-Adjustment Methods

Young (1990a) proposed an inventive use of Item Response Theory (IRT), in which course grades are treated as item scores and estimated theta becomes the intended adjusted achievement index. Young (1990a) first factor analyzed the course grades and separated the courses into relatively unidimensional groups based on the factor analysis results, then applied a restricted version of Samejima's (1969) Graded Response Model on the separated groups of courses. He found that the predictability of the IRT-based GPA from SAT-V, SAT-M, and high school GPA exceeded that of the unadjusted GPA. The increase in squared multiple correlation (R^2) ranged from .0015 to .0955, from negligible to sizable, depending on possibly a number of factors which included the distribution of grades, number of courses involved in the adjustment, number of courses taken by each student, or even the clarity of the defining trait (Young, 1990a).

Moreover, Young (1990a) advocated that other polytomous IRT models may be applicable to grade adjustment. However, only Samejima's (1969) graded response model in the polytomous IRT family has been examined. Therefore, this study intends to expand the IRT-based category applied to grade adjustment. Models with different levels of complexity in terms of number of model constraints or, alternatively, number of free model parameters (model parameters to be estimated from the data) involved are selected. The rating scale model (Andrich, 1978) has the fewest free parameters (the most model constraints), followed by the partial credit model (Masters, 1982), and then the generalized partial credit model (Muraki, 1992) and the graded response model (Samejima, 1969). The latter two models have the same number of free parameters (see Appendix A for a list of the models used for adjusting grades).

Suppose there are k distinct grades within each course, and that there are m courses. The rating scale model has k-1 step parameters (steps between grades are constrained to be the same across courses) and m location or difficulty parameters, one for each course. In the partial credit model, the steps between grades are allowed to vary across courses, and the location parameters are embedded in the steps for different courses; therefore, it has $m(k\text{-}1)$ parameters. Neither the rating scale nor the partial credit model has course discrimination parameters, or implicitly, the course discrimination parameters are the same across courses and equal to 1 (the so-called polytomous Rasch models). The generalized partial credit model and the graded response model, on the other hand, have m course discrimination parameters (slopes), one for each course, in addition to $m(k\text{-}1)$ "step" parameters (intercepts).

The IRT models aforementioned allow the distances or steps between grades assigned within courses to vary. The models proposed by Caulkins, Larkey, and Wei (1996), however, constrained the steps between grades to be the same within classes. The additive, multiplicative and combined methods provided by Caulkins, et al. can be respectively represented by the equations: $x_{ij} = \theta_i - \beta_j + \varepsilon_{ij}$, ($i$=1,…,$N$; j=1,…,m); $x_{ij} = \theta_i / \alpha_j + \varepsilon_{ij}$, ($i$=1,…,$N$; j=1,…,m); and $x_{ij} = \theta_i / \alpha_j - \beta_j + \varepsilon_{ij}$, ($i$=1,…,$N$; j=1,…,m), where x_{ij} represents the observed grade for person i in course j, θ_i is the person parameter or the adjusted GPA, α_j and β_j are course parameters representing course difficulties or grading standards, and ε_{ij} is the error term. The α_j parameter controls the magnitude of the grade steps across courses, while the β_j parameter adjusts the location of the courses. Altogether, there are m course parameters for the additive and multiplicative models, and $2m$ parameters for the combined model. The Caulkins, et al. models are referred to as linear models because the adjusted GPA using these models has a linear relationship to the unadjusted GPA. The authors listed least-squares solutions (minimizing sum

of squared error over persons and courses) for the person and course parameters and the unknown parameters can be estimated through simple iterative procedures with reasonable initial estimates.

Caulkins, et al. (1996) tested their models with a cohort at Carnegie Mellon University and found that adjusted GPAs had higher correlations than did unadjusted GPA with high school GPA, SAT-M, SAT-V, and high school rank. The authors also compared their additive method with the IRT based method suggested by Young (1990a) and found that the predictability of both IRT and linearly adjusted GPAs by high school GPA, SAT-M, and SAT-V improved over unadjusted GPA by the same predictors. Furthermore, the increase in R^2 was larger for the additive model than that for the IRT-based method. Therefore, the simpler linear models with fewer free parameters (when the number of distinct grades involved in adjustment is larger than 2) may be superior to the relatively more complicated IRT method in adjusting GPA.

However, research on the two categories of grade adjustment methods, IRT-based and linear, has been exploratory. Linn (1966) called for studies of grade adjustment methods to cross-validate results. His call, however, was largely ignored or forgotten by the studies conducted after his review. Therefore, previous findings about grade adjustment did not take into account same-sample effects. Had same-sample effects been corrected, actual improvement in R^2 and relative performance of the methods may have been different. The current study attempted to incorporate the cross-validation criterion in addition to the multiple correlation criterion (with a standardized test, high school GPA, and/or high school rank) commonly used in grade adjustment studies. Results of one cohort will be cross-validated from course parameters estimated from a previous cohort.

The performance of grade-adjustment methods over successive years or cohorts is an important practical issue. Operational use of adjusted-GPA measures could well involve applying estimates of course difficulty parameters derived from one cohort to the grade data of a later cohort. Moreover, regression weights derived from a previous cohort may be used to predict an incoming cohort's college achievement for admission or scholarship selection purposes. Therefore, in addition to cross-validating the course parameter estimates, prediction accuracy of the different adjusted-GPA measures using regression weights of the preadmission variables derived from a previous cohort will also be examined.

Furthermore, because consistency of students' class rank with their relative standings on the "achievement" continuum is a desired goal, internal order consistency (IOC) rate is also included as a criterion. IOC rate, defined as the proportion of times in which two students taking the same course received distinct grades in the same order as their achievement index, represents a form of the fit of the model to the data. Let "Tot" represent the total number of cases in which two students received non-tied grades in the same course. This number includes all possible pairs of students and does not depend on any achievement measure. For each achievement measure, "Hits" is number of pairs (of students who received different grades in the same course) that were not tied on the given achievement measure, and for which the achievement measure correctly predicted which student received the higher grade. Algebraically, IOC rate is the quotient of Hits/Tot.

In summary, the purpose of this study is threefold. First, it expands the IRT-based category used in grade adjustment. Second, it answers the long neglected call made by Linn (1966) by correcting for same-sample effects to provide a more realistic evaluation of the adjustment methods. Third, it evaluates prediction precision of the adjusted achievement

measures using preadmission variables, an important practical use of college achievement measure. It is hoped that by reducing errors associated with different grading practices and different course taking patterns, adjusted GPAs will be more comparable among students. This study attempts to evaluate these alternative achievement measures under operationally realistic conditions to substantiate the advantage, or the lack thereof, of using these alternative measures. In response to the recent debate about grade inflation, this study may provide timely information to administrators or policy decision-makers.

Study 1

Method

Source of Data

The data consisted of grades of college freshmen in 21 courses over two consecutive years, 1995/96 and 1996/97 ('95 and '96 respectively for brevity). Course grades were reported on an A to F scale and were coded as A=4 and F=0. These data were not complete in terms of representing all of the courses taken by college first-year students, or all of the courses taken by any given college first-year student. Moreover, the college underwent some unusual changes in the enrollment requirements for the two cohorts. Therefore, this study is considered exploratory and will be replicated with other data more typical of college grades. However, they represent incomplete ordinal data with substantial variation in course difficulty and course-taking patterns. A total of 1,255 students in the '95 cohort and 1,796 students in the '96 cohort are represented by at least one grade.

In addition to college grade data, ACT Assessment scores and self-reported grades in high school courses were available for 1,710 of the 1,796 students in the '96 cohort. The composite score on the ACT Assessment was used as the admissions test score. Self-reported

high school grades were taken from the Course Grade Information Section of the ACT Assessment. This section includes 30 high school courses representing standard college preparatory courses such as English I to IV, advanced courses such as calculus, and ancillary coursework such as art, music, and foreign languages.

Procedures

College course grade data from the '95 and '96 cohorts were separately fitted with 4 models using Multilog v.6 (Thissen, 1991): 1) partial credit model with a common slope parameter *a* estimated from the data (PCMA), 2) partial credit model with a common slope parameter *a* fixed to 1 (PCM1), 3) generalized partial credit model with all slope parameters free to vary (GPCM), and 4) graded response model (GRM). It may be worthwhile to note that all partial credit models are special cases of Bock's (1972) Nominal model (Thissen & Steinberg, 1986). For model set-up of these models in Multilog, see Multilog manual (Thissen, 1991) and Thissen and Steinberg (1986). The model set-up for the rating scale model in Multilog is currently unavailable.

The same data were fitted with the rating scale model (RSBS) using Bigsteps (Wright & Linacre, 1990). To check whether different constraints and estimation methods used in different programs may make a difference, the partial credit model analysis was repeated with Bigsteps (PCBS). However, the Bigsteps program does not perform the generalized partial credit model and the graded response model. Therefore, results of 6 IRT models were available for further analyses, three of which are essentially the same models with only differences in how parameters were estimated or how the models were identified. For a perusal of the different polytomous IRT models, see van der Linden and Hambleton (1997). In addition, simple iterative procedures based

on Caulkins, et al. (1996) solutions were used to estimate the person and course parameters for the three linear models: additive, multiplicative and combined.

Person parameters for the '96 cohort were estimated twice, first with course parameters estimated from the same '96 sample (same-sample achievement indices), and then cross-validated with course parameters estimated from the '95 sample (cross-validated achievement indices). Therefore, the effect of same-sample effects of the course parameter estimates was examined with all 6 IRT models and 3 linear models for exploratory purpose. Because these data were not typical of freshman college grades and regression weights derived from such uncharacteristic data would not be ideal for operational use, the cross-validity of regression weights for predicting the achievement measures by the preadmission variables was not examined in this study.

Multiple correlations of the same-sample and cross-validated college achievement indices (unadjusted and adjusted college GPAs) with ACT composite score and high school GPA were computed. Pairwise comparisons of correlations were performed using the Z_2^* statistic documented by Steiger (1980). The Z_2^* statistic was appropriate for comparing correlations between r_{jk} and r_{hm} and was therefore applied to multiple correlations of the college achievement indices (variables j and h) with their predicted scores from the preadmission variables (variables k and m). Given the large number of tests performed, multistage Bonferroni correction recommended by Larzelere and Mulaik (1977) were used to maintain the nominal type I error rate of .05. In addition, IOC rates, defined as the proportion of times in which two students taking the same course received non-tied grades in the same order as their achievement indices (i.e., higher course grade with higher achievement index and vice versa), were tabulated for the different college achievement indices.

Results

Results reported in this session are pertaining to the '96 cohort. Summary statistics for same-sample and cross-validated college achievement measures are reported in Table 1. Raw GPA and GPAs adjusted by different methods were placed on different scales, so the standard deviations and ranges were not comparable. It is probably worth noting that the Multilog and Bigsteps programs centered the person parameters at different locations and put them on different scales for the supposedly equivalent partial credit models.

TABLE 1

Summary Statistics for Same-Sample and Cross-validated College Achievement Measures of the '96 Cohort (Study 1)

College achievement measures	Same-sample				Cross-validated			
	Mean	SD	Min	Max	Mean	SD	Min	Max
GPA	2.51	0.97	0.00	4.00	-	-	-	-
Additive	2.52	0.99	-0.59	5.37	2.66	0.98	-0.43	5.06
Multiplicative	2.56	1.02	0.00	6.57	2.65	1.05	0.00	5.84
Combined	2.44	1.18	-4.88	5.18	2.68	0.94	-4.70	4.67
RSBS	0.30	1.98	-5.18	6.04	0.73	2.17	-5.75	6.80
PCBS	0.30	2.11	-5.92	6.70	0.80	2.48	-6.81	7.81
PCM1	-0.02	0.81	-2.50	2.39	0.04	0.87	-2.92	2.57
PCMA	-0.02	0.81	-2.42	2.33	0.05	0.88	-3.10	2.49
GPCM	-0.02	0.81	-2.49	2.17	0.05	0.89	-2.85	2.34
GRM	0.00	0.83	-2.62	2.20	0.03	1.07	-5.12	2.31

Note. Same-sample achievement measures are based on course parameters estimated from the '96 cohort. Cross-validated achievement measures are based on course parameters estimated from '95 cohort. IRT model abbreviations: RSBS=rating scale performed by Bigsteps, PCBS=partial credit performed by Bigsteps, PCM1=partial credit with common slope of 1 performed by Multilog, PCMA= partial credit with common slope estimated from the data performed by Multilog, GPCM=generalized partial credit performed by Multilog, and GRM=grade response performed by Multilog. N=1,710.

Multiple correlations of college achievement measures with preadmission variables (ACT composite score and high school GPA) and IOC rates of the achievement measures with college course grades are presented in Table 2. As expected, multiple correlations and IOC rates of same-sample adjusted college GPAs (course parameters estimated from the '96 cohort) were higher than their cross-validated counterparts (course parameters estimated from the '95 cohort). Pairwise comparisons of the multiple correlations were performed within the same-sample and cross-validated set of college achievement measures (see first two columns of Table 2).

TABLE 2

Multiple Correlations with Preadmission Variables and Internal Order Consistency Rates with Course Grades for College Achievement Measures of the '96 Cohort (Study 1)

College achievement measures	Multiple correlation with ACT composite and H.S. GPA		Internal order consistency rate	
	Same-sample	Cross-validated	Same-sample	Cross-validated
GPA	**.417$_a$**	**.417$_a$**	**.831**	-
Additive	.543$_c$	.445$_{bc}$	.868	.850
Multiplicative	.516$_b$	.443$_{abc}$	**.873**	.850
Combined	.533$_{bc}$	.443$_{abc}$	.855	.850
RSBS	**.568$_d$**	**.462$_c$**	.869	.847
PCBS	.566$_d$	.451$_{bc}$	.872	.847
PCM1	.532$_{bc}$	.435$_{ab}$	.869	.849
PCMA	.537$_{bc}$	.440$_{ab}$	.870	**.851**
GPCM	.532$_{bc}$	.451$_{bc}$	.860	.846
GRM	.542$_{bc}$	.409$_{ab}$	.860	.836

Note. Correlations in the same column that do not share subscripts differ at $p<.05$ (with multi-stage Bonferroni correction) in the Steiger (1980) test of significant correlation difference. Same-sample achievement measures are based on course parameters estimated from the '96 cohort. Cross-validated achievement measures are based on course parameters estimated from '95 cohort. GPAs are not crossed-validated; they are listed under Cross-validated only for comparison with crossed-validated achievement measures. IRT model abbreviations: RSBS=rating scale performed by Bigsteps, PCBS=partial credit performed by Bigsteps, PCM1=partial credit with common slope of 1 performed by Multilog, PCMA= partial credit with common slope estimated from the data performed by Multilog, GPCM=generalized partial credit performed by Multilog, and GRM=grade response performed by Multilog. N=1,710.

With same-sample bias, in terms of multiple correlations with preadmission variables, all adjusted GPAs (ranging from .516 to .568) performed significantly better than the unadjusted GPA (.417). Moreover, the additive model appeared to be the best among the linear models (.543 vs. .516 and .533), although the difference between the additive and combined models was not statistically significant. The two Bigsteps models excelled among the IRT models (.568 and .566 vs. .532 to .542 for other IRT models). The Multilog models did not differ significantly from each other or from the linear models. When same-sample bias was taken into account, however, only the additive (.445), RSBS (.462), PCBS (.451), and GPCM (.451) models improved significantly over raw GPA (.417) on multiple correlation with ACT composite score and high school GPA. Regardless of whether same-sample bias existed, the two Bigsteps models appeared to perform the best among all the college achievement measures included in this study.

In terms of IOC rates with course grades, adjusted GPAs (ranging from .855 to .873 for same-sample and from .836 to .851 at cross-validation) still performed better than the unadjusted GPA (.831) despite the correction of same-sample bias (see last two columns of Table 2). The multiplicative model had the highest IOC rate (.873) with same-sample bias, but the PCMA model had the best IOC rate (.851) at cross-validation. Since the differences in IOC rates were so minor, perhaps we should not put too much weight on this criterion in informing the relative performance of the adjustment models.

Discussion

In spite of the problematic characteristics of the college grade data used in this study, the advantage of the alternative measures to GPA seemed fairly clear. Any adjustment of college GPA was better than no adjustment. This observation was true on both criterion measures and remained true at cross-validation. Since differences on IOC rates could not be tested for

statistical significance, the relative performance of the models observed on this criterion might have been due to chance, and should not be relied on too heavily for model selection. Because the relative performance of the models on multiple correlations was tested for statistical significance, it seemed to be reasonable to conclude that the Bigsteps IRT models, especially the rating scale model, outperformed the other models. The improvement of RSBS adjusted GPA over the unadjusted GPA on predictive validity was quite substantial. The squared multiple correlation with ACT composite score and high school GPA increased by .149, from $(.417)^2$ to $(.568)^2$, with same-sample effects. Although the cross-validated results may have been underestimated due to uncharacteristic changes between the two cohorts analyzed in this study, the improvement remained sizable at cross-validation, R^2 raised by .04 from $(.417)^2$ to $(.462)^2$.

Moreover, this study showed the importance of cross-validation envisioned by Linn (1966). The relative performance of the different models did change depending on whether same-sample bias was taken into account. For instance, with same-sample bias the better performing GRM-adjusted GPA in relation to simple GPA (.542 vs. .417) on multiple correlation became worse than simple GPA (.409 vs. .417) at cross-validation, though the difference was not statistically significant. Moreover, the relative standings of the adjustment methods, especially the top performing ones, changed on IOC rate, depending on whether same-sample bias was considered. One may wonder if these inconsistencies were due to chance or to the peculiarities of the data used for cross-validation in this study. In search of a solution to this uncertainty, Study 2 will employ another data set more typical of college course grades.

It is interesting to note that the supposedly equivalent partial credit model fitted by the Bigsteps and Multilog programs performed differently on multiple correlation with the preadmission variables. The advantage of the Bigsteps partial credit model over the Multilog

counterpart on multiple correlations may be due to fewer distributional constraints on the theta scale (without vs. with normal prior), different estimation procedures (joint maximum likelihood vs. marginal maximum likelihood), or the linearization of the achievement scale made by the Bigsteps program. Regardless of the existence of same-sample bias, the IRT models fitted by the Multilog program (PCM1, PCMA, GPCM, and GRM) did not perform differently from each other. Similarly, the models fitted by the Bigsteps program (RSBS and PCBS) did not perform differently from each other. Therefore, the differences in performance among the IRT models appeared to be attributable to the different estimation characteristics employed by the programs rather than to model complexity in terms of the number of free parameters.

When Multilog is used to fit IRT models, it should be cautioned that "Multilog does not check for zero-category frequencies, the erroneous results that are obtained if a category has no respondents are unpredictable" (Thissen, 2000). When this message is translated in terms of grade data, Multilog should not be used if there are missing course grades in any class unless intercepts surrounding the missing grades are not estimated. This can be accomplished by collapsing grades or by fixing the intercept values around the missing categories. Because there were some missing grades in the '95 data, cross-validated results based on Multilog models may have been distorted and should be interpreted with caution. Bigsteps, on the other hand, allows the option of keeping or collapsing the missing grade. It is not clear, however, how missing grades would affect the stability of their step estimates and how they would affect the cross-validation results in this study. Therefore, Study 2 will consider fixing the step parameters (for the Bigsteps program) or the intercepts (for the Multilog program) surrounding the missing grades.

Due to the peculiarities of the data, findings of this study should be viewed as tentative, especially on the parts pertaining to cross-validation. Results of this study should be replicated with college grade data consisting of characteristics that are representative of the college. If alternative measures to GPA were to become operational, course parameters and regression weights would likely be estimated for individual colleges from a representative cohort. Then, these course parameter estimates may be used for GPA adjustment and the regression weights for prediction of achievement of later cohorts. Study 2 attempts to evaluate the merits of adjusting GPA under these realistic practical conditions with data on typical courses offered by colleges.

Study 2

Method

Students

Students were drawn from two consecutive cohorts at a large midwestern university. For each cohort, an initial pool of students was created by matching ACT high school profile records with course grade records from this university. ACT high school profile records are organized by year of graduation from high school. The cohorts graduated from high school in 1995 and 1996. The initial pool for each cohort included students who earned a letter grade (F to A+) in at least one course in each of the fall and spring semesters of the second academic year following their high school graduation. It was necessary to use second-year grades because first-year grade data were inadvertently missing from the course grade file. [The use of second-year grade data could conceivably underestimate the predictive validity of preadmissions tests such as the ACT Assessment, which is intended to predict first-year college achievement, but second-year data was expected to be adequate for the main purpose of this study, which was to compare models under both same-sample and cross-validity conditions. See discussion section for further

information on this point.] There was attrition from these pools because students had to have at least one letter grade in the courses included in this study, and not all courses were included (see below). The final numbers of students in the '95 and '96 cohorts (so named by year of high school graduation) were 1,823 and 1,879 respectively. The number of grades per student ranged from1 to 12, with 95% of students represented by 3 to 9 grades.

College courses

As indicated above, the courses in this study were taken during each cohort's second year at the University of Iowa. Two types of courses were defined: individual courses and catch-alls. Individual courses include any course in which thirty or more students from each cohort earned a letter grade. Since individual courses are one semester in duration, but are sometimes offered both fall and spring semesters, grades were pooled from both semesters, if applicable. Individual courses are designated by a department code and a course code.

Catch-alls represent one or more smaller courses within the same department. A similar procedure of pooling grades from small courses within departments was used by Stricker, et al., (1994) and Elliot and Strenta (1988). Courses subsumed within catch-alls had fewer than thirty letter grades for one or both of the cohorts. A catch-all might not represent the same mixture of individual courses in both cohorts. In order to make the catch-alls for a department more alike across cohorts, they were defined separately by semester. Like individual courses, a catch-all was required to have thirty or more letter grades within each cohort. With these procedures, a department could be represented by a catch-all for each semester, for just one semester, or by no catch-all. Catch-alls are designated by a department code and a semester code.

In total, there were 146 courses of both types: 94 individual courses and 52 catch-alls. These courses represented 85% and 83% of the letter grades received by the 1995 and 1996

cohorts respectively. The individual courses represented 41 departments. The catch-all courses represented 32. The total number of departments represented was 53.

College grade coding and GPA

The letter-component of each grade was numerically coded from 0 (F) to 4 (A). This coding was used to compute the GPA as well as to compute all adjusted-GPA measures. The pluses and minuses that are attached to letter grades at the University of Iowa were not used. The numerical grades were not weighted by number of credit hours. These procedures are similar to those used in other GPA-adjustment studies (e.g., Young, 1990a; Stricker, et al., 1994; Caulkins, et al., 1996; Johnson, 1997).

Procedure

The basic model fitting and cross-validation procedures used in this study were similar to Study 1. Several modifications were made based on the findings and discussions of Study 1. First, because the partial credit model fitted with Bigsteps performed similarly on IOC rate and better on multiple correlation with preadmission variables in comparison to the Multilog counterparts, the partial credit model was not re-fitted with the Multilog program.

Second, to control for possible idiosyncratic effects of different ways of handling missing grades in different polytomous IRT programs, missing grades within courses were handled differently from Study 1. Instead of using the default methods of handling missing categories, step parameters surrounding the missing grade within any course for the PC model were fixed to the corresponding common step parameters estimated for the RS model. Similarly, intercepts surrounding the missing grade within any course for the GR or GPC models were fixed to the corresponding common intercept values obtained from a previous analysis on the same data by constraining the intercepts to be the same across courses.

Third, in addition to cross-validating course parameter estimates, equations for predicting college achievement measures from the preadmission variables were also cross-validated. Regression equations were derived from the '95 cohort for the college achievement measures. Predicted achievement scores of the '96 cohort were obtained by substituting their ACT composite scores and reported high school GPAs into the regression equations derived from the '95 cohort. Scatter plots of the same-sample and cross-validated college achievement measures of the '96 cohort against the corresponding predicted achievement scores were obtained. Correlations of these bivariate plots were computed and tested for statistical significant difference among the achievement measures using the Z_2^* statistic proposed by Steiger (1980).

Finally, three IOC rates were computed for each achievement measure. To represent these rates formulaically, let "Tot" represent the total number of cases in which two students received non-tied grades in the same course. This number includes all possible pairs of students and does not depend on any achievement measure. For each achievement measure, two more numbers were obtained: "Ties" is the number pairs (of students who received different grades in the same course) that were tied on the given achievement measure, and "Hits" is number of pairs that were not tied on the given achievement measure, and for which the achievement measure correctly predicted which student received the higher grade. The three IOC rates computed for each achievement measure were 1) Hits/Tot, 2) Hits/(Tot – Ties), and 3) (Hits + 0.5*Ties)/Tot. In rate 1, ties are treated as prediction errors (ties = errors). This was the rate computed in Study 1. In rate 2, ties are treated as missing (ties = missing). In rate 3, ties are treated as contributing to a 0.5 IOC rate (ties = half hits). The effect of these different ways of handling ties in achievement measures was examined.

Results

Regression weights and their estimated standard errors derived from the '95 cohort for predicting college achievement measures from preadmission variables are listed in Table 3. Students' ACT composite scores and high school GPAs (conditioned on each other) were significant predictors ($p < .0001$) for all of the college achievement measures included in the study. Squared multiple correlations were higher for the adjusted GPA equations (ranging from .231 to .259) than for the raw GPA equation (.162), but the differences in R^2 were less than .1.

TABLE 3

Beta Weights Estimated from the '95 Cohort for Prediction of the '96 Cohort (Study 2)

College achievement measure	Intercept	ACT Composite	High school GPA	R^2
GPA	.750 (.121)	.019 (.004)	.499 (.034)	.162
Additive	.093 (.120)	.031 (.004)	.604 (.034)	.245
Multiplicative	.248 (.119)	.028 (.004)	.581 (.034)	.231
Combined	-.302 (.137)	.035 (.004)	.679 (.039)	.241
RS	-6.972 (.377)	.108 (.011)	1.920 (.107)	.259
PC	-7.032 (.376)	.108 (.011)	1.905 (.107)	.257
GR	-3.790 (.163)	.043 (.005)	.809 (.046)	.243
GPC	-3.833 (.165)	.045 (.005)	.842 (.047)	.254

Note. Values in parentheses are standard errors of the beta estimates. IRT model abbreviations: RS=rating scale, PC=partial credit, GR=graded response, GPC= generalized partial credit. N=1,789.

Summary statistics for predicted, same-sample, and cross-validated achievement measures of the '96 cohort are given in Table 4. Predicted achievement measures were computed by substituting '96 students' ACT composite scores and high school GPAs into the prediction equations derived from the '95 cohort (see Table 3). Same-sample achievement measures were based on course parameters estimated from the same '96 cohort, while the cross-validated

counterparts were based on course parameters estimated from the '95 cohort. Because there were missing data on the preadmission variables, only 1,822 of the 1,879 students had predicted achievement measures. Means and standard deviations of the same-sample and cross-validated achievement measures were very close to each other. Means of the predicted achievement measures were higher than means of the same-sample or cross-validated achievement measures. As a group, the '96 cohort achieved somewhat less than expected. Standard deviations of the predicted measures, however, were much smaller than the same-sample or cross-validated counterparts. This makes sense because the prediction equations were derived using least-squares solutions.

TABLE 4

Summary Statistics for College Achievement Measures of the '96 Cohort (Study 2)

College achievement measures	Predicted (*N*=1,822)		Same-sample (*N*=1,879)		Cross-validated (*N*=1,879)	
	Mean	SD	Mean	SD	Mean	SD
GPA	2.98	0.24	2.95	0.62	-	-
Additive	2.98	0.31	2.94	0.65	2.95	0.65
Multiplicative	2.99	0.30	2.95	0.64	2.96	0.64
Combined	2.95	0.35	2.92	0.71	2.94	0.71
RS	2.44	1.02	2.35	2.03	2.37	2.05
PC	2.32	1.01	2.21	2.04	2.25	2.04
GR	0.12	0.42	0.08	0.85	0.09	0.89
GPC	0.23	0.44	0.03	0.88	0.21	0.88

Note. Predicted achievement measures are computed by substituting '96 students' ACT composite scores and their high school GPAs into the prediction equations derived from the '95 cohort. Same-sample achievement measures are based on course parameters estimated from the '96 cohort. Cross-validated achievement measures are based on course parameters estimated from the '95 cohort. IRT model abbreviations: RS=rating scale, PC=partial credit, GR=graded response, GPC= generalized partial credit.

The correctness of predicted college achievement measures from the preadmission variables is evaluated through scatter plots of the observed achievement measures against their predicted counterparts and correlations of those bivariate plots. These plots are shown in Appendix A. The scatter plot of the baseline unadjusted GPA against predicted GPA is given in Figure 1 (the upper left graph). Plots for sample-sample adjusted achievement measures are presented in Figures 1 (except the upper left graph) and 2, and plots for cross-validated adjusted achievement measures are presented in Figures 3 and 4. Correlations of the plots are given in the first two columns of Table 5. Although the correlations were moderate (around .50s), the majority of the points were within the 95% confidence intervals of the regression lines. It appeared that there were more points below the lower interval than points above the upper interval for the unadjusted and linearly adjusted GPAs, while the points distributed quite evenly outside the upper and lower intervals for the IRT adjusted GPAs. This difference may be because the raw GPA and the linearly adjusted GPAs were bounded and had some ceiling effect while the IRT adjusted GPAs were, theoretically, unbounded.

The last two columns of Table 5 provide multiple correlations of the achievement measures with the ACT composite score and high school GPA using regression weights derived from the same '96 cohort. Multiple correlations of the achievement measures with the preadmission variables using equations estimated from '95 were similar to the corresponding multiple correlations using equations estimated from '96. Moreover, regardless of the year of equations used, the same-sample multiple correlations were, as expected, larger than their cross-validated counterparts, but only by a very narrow margin. The beta weights and course parameter estimates appeared to be stable over the two cohorts.

TABLE 5

Multiple Correlations of College Achievement Measures with Preadmission Variables for the '96 Cohort (Study 2)

College achievement measures	'95 equations		'96 equations	
	Same-sample	Cross-validated	Same-sample	Cross-validated
GPA	**.411$_a$**	**.411$_a$**	**.413$_a$**	**.413$_a$**
Additive	.497$_c$	.492$_c$	.498$_c$	.493$_c$
Multiplicative	.484$_b$	.478$_b$	.486$_b$	.480$_b$
Combined	.504$_{cd}$	.497$_{cd}$	.505$_{cd}$	.499$_{cd}$
RS	.517$_d$	.511$_d$	.518$_d$	.513$_d$
PC	.516$_d$	.508$_d$	.517$_d$	.509$_d$
GR	.505$_{cd}$	.487$_{bc}$	.506$_{cd}$	.490$_{bc}$
GPC	.508$_{cd}$	.499$_{cd}$	.509$_{cd}$	.501$_{cd}$

Note. Correlations in the same column that do not share subscripts differ at $p<.05$ (with multi-stage Bonferroni correction) in the Steiger (1980) test of significant correlation difference. Same-sample achievement measures are based on course parameters estimated from the '96 cohort. Cross-validated achievement measures are based on course parameters estimated from the '95 cohort. GPAs are not crossed-validated; they are listed under cross-validated only for comparison with cross-validated achievement measures. IRT model abbreviations: RS=rating scale, PC=partial credit, GR=graded response, GPC= generalized partial credit. N=1,822.

The multiple correlations of all types of adjusted-GPA (by any model) with preadmissions variables were significantly improved over those for raw GPA (see Table 5). R^2 increments (within column differences between squared values of Table 5 and squared values of the baseline GPA in the first row) ranged from .065 to .097 for same-sample and from .059 to .092 for cross-validated achievement measures, and the ranges were nearly the same regardless of the year of prediction equations used. The RS method consistently outperformed the other adjustment methods on multiple correlations with the preadmission variables, although the performance of the combined method in the linear category and the PC and GPC methods in the IRT category was not significantly different from the RS method.

TABLE 6

Internal Order Consistency Rates of College Achievement Measures with Course Grades for the '96 Cohort (Study 2)

College achievement measures	Ties = errors		Ties = missing		Ties = half hits	
	Same-sample	Cross-validated	Same-sample	Cross-validated	Same-sample	Cross-validated
GPA	.837	-	.852	-	.846	-
Additive	.858	.856	.858	.856	.858	.856
Multiplicative	.858	.856	.858	.856	.858	.856
Combined	.853	.852	.853	.852	.853	.852
RS	.858	.856	.859	.857	.859	.857
PC	.859	.856	.860	.857	.860	.856
GR	.855	.848	.855	.849	.855	.849
GPC	.854	.848	.854	.848	.854	.848

Note. Same-sample achievement measures are based on course parameters estimated from the '96 cohort. Cross-validated achievement measures are based on course parameters estimated from the '95 cohort. IRT model abbreviations: RS=rating scale, PC=partial credit, GR=graded response, GPC= generalized partial credit. *N*=1,879.

The relative performance of the adjustment methods on different definitions of IOC rates is shown in Table 6. How ties in achievement measures were treated mattered little to the adjusted GPAs, but it seemed to affect raw GPA quite substantially because raw GPA had more ties than the adjusted GPAs. Raw GPA was disadvantaged if ties were treated as errors but benefited the most if ties were treated as missing; perhaps it is fairer to treat ties as half hits and half errors. Regardless of how ties in achievement measures were treated, raw GPA (ranging from .837 with ties=errors to .852 with ties=missing) was inferior to most adjusted GPAs in terms of IOC rates except for the cross-validated GR (.849) and GPC (.848) measures when ties were treated as missing. However, the differences in IOC rates appeared to be small (not very discriminating) and the differences were not conducive to tests of statistical significance. Moreover, it is debatable conceptually how ties should be treated. Therefore, the IOC rate

BEST COPY AVAILABLE

criteria do not seem to be very informative on the relative performance of the adjustment methods.

General Discussion

The primary purposes of this research report are to expand the IRT-based category used in grade adjustment, and correct for same-sample bias to provide a more realistic picture of actual implementation. These objectives are achieved with course grade data and ACT test reports from consecutive cohorts of two universities, results of which are reported in two studies.

Several consistent findings are observed over the two studies. First, any adjustment of GPA is better than no adjustment in terms of multiple correlations with the preadmission variables and, for the most part, on IOC rates as well. This observation is true for both the same-sample and cross-validated sets of alternative achievement measures. Second, among the adjustment models included in this report, the rating scale and partial credit models fitted by the Bigsteps program excelled in terms of multiple correlations with the preadmission variables. Third, the graded response model, on the other hand, is the most unstable among the adjustment methods across cohorts on multiple correlations with the preadmission variables (it has the largest drop in R^2 at cross-validation, by .126 in Study 1 and by .016 in Study 2). Fourth, the IOC rate is not very informative about the relative performance of the adjustment methods because the observed differences among the models are very small and they are not conducive to statistical tests.

The magnitude of same-sample improvement in R^2 of adjusted over unadjusted GPA (ranging from .092 to .149 in Study 1, and from .065 to .097 in Study 2) is, in general, similar to other grade adjustment studies reported in the literature. In particular, R^2 increments for the graded response model were .120 in Study 1 and .085 in Study 2, while those reported by Young

(1990b) and Stricker, et al. (1994) were respectively .073 and .064. R^2 increments for the additive model were .121 in Study 1 and .077 in Study 2 while those reported by Caulkins, et al. (1996) and Johnson (1997) were respectively .111 and .086. The similarity of R^2 increments is particularly noteworthy given across-study differences in the selectivity of institutions represented, predictor variables, and in time frame represented by the college achievement measures. Caulkins et al., (1996), Johnson (1997), and Stricker (1994) all used the SAT rather than the ACT Assessment to predict college achievement measures. Stricker (1994) used high school rank rather than high school GPA. Caulkins, et al., (1996) and Johnson (1997) used data from highly selective institutions, whereas Stricker, et al. (1994) used data from a less selective, public university. Johnson (1997) used four-year cummulative grades, Caulkins et al., (1996) used three-year cumulative grades, and Stricker et al., (1994) used first-year grades exclusively.

Cross-validated results (of course parameters and regression weights) are more reflective of operational use of the adjustment methods and should be examined in all grade adjustment studies (Linn, 1966). As shown by the two studies in this report, cross-validated results can vary quite substantially from institution to institution. R^2 differences between same-sample and cross-validated indices ranged from .070 to .126 in Study 1 and from .005 to .016 in Study 2. As mentioned before, the cross-validated results of Study 1 may have been distorted. It is not clear, however, if the small changes at cross-validation in Study 2 are characteristic of other institutions. This matter should be examined in future studies.

Moreover, stability of the adjustment methods over consecutive cohorts may be an important consideration for model selection. Then, guided by the results of Study 2, the linear models, especially the additive model, and the rating scale model have the smallest drop in multiple correlation with the preadmission variables at cross-validation. This makes sense

because the linear models and the rating scale model have fewer free parameters to be estimated than the other adjustment models. Perhaps with the exception of the graded response model (its R^2 change was .016 and the largest), the differences among the adjustment methods in R^2 decrease at cross-validation were quite small (R^2 change ranged from .005 to .008) and could have been due to chance errors. Therefore, the relative stability of the adjustment methods observed in this study should be replicated in future studies.

In the IRT-based family, only Samejima's (1969) graded response model has been applied to grade adjustment. This report expanded the list to include other appropriate polytomous IRT models as well as to investigate their implementation with different programs. Results showed that the graded response model is not the best IRT-based model for grade adjustment, particularly in terms of stability over cohorts. In addition, the differential performance of the IRT models may not be due to model complexity per se, but to the program and the estimation procedures employed therein that carried out the modeling.

All of the models included in this report assume unidimensionality of course grades assigned to students, implicitly or explicitly. Young (1990a) factor-analyzed course grades before applying the IRT graded response model to separate sets of relatively homogeneous courses. In this report, the IRT models were applied to the whole set of courses available without testing the dimensionality of the courses. It is unlikely that the courses would be unidimensional. It is not clear how the violation of the unidimensionality assumption has affected the results of the IRT models reported here. However, the effect of violating the dimensionality assumption is not necessarily unique to the IRT models, because the raw GPA and the linearly adjusted GPAs also implicitly treat course grades as unidimensional. If one is interested in multidimensional

measures of achievement, perhaps profiles of achievement scores on unidimensional sets of courses should be examined rather than a single achievement measure.

We can think of no reason why the relative performance of the models in this study, under both same-sample and cross-validity conditions, should depend on the fact that achievement was measured by second-year grades exclusively. GPA-adjustment studies have used four-year cummulative grade data (Young, 1990a; Johnson, 1997), three-year cummulative grade data (Caulkins, et al., 1996), first-year grades (Stricker, et al., 1994), and other single-year grade data including second-year grades exclusively (Caulkins, et al., 1996). It is conceivable that the correlation of any measure based on second-year grades exclusively could underestimate the predictive validity of preadmissions tests like the ACT Assessment, which are intended to predict first year achievement. However, the relative performance of GPA-adjustment methods has not been attributed to the particular time frame of achievement represented in studies, (e.g., Johnson, 1997; Caulkins, 1996; and Stricker, et al., 1994).

In terms of model selection, based on the results of the two studies, it seems to be reasonable to recommend the rating scale model or the partial credit model performed by the Bigsteps program, especially if high multiple correlation of the achievement measure with the preadmission variables is desired. On the other hand, model simplicity and ease of implementation may be important considerations for operational use of alternative measures, then the linear models, particular the additive model, would perform sufficiently well. It is a matter of judgment, however, if the magnitude of improvement in the predictive validity coefficients and IOC rates with course grades observed in the two studies of this report warrant the controversial operation of adjusting the ordinary GPA. This administrative choice is left for the readers to ponder upon.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bejar, I. I., & Blew, E. O. (1981). *Grade inflation and the validity of the Scholastic Aptitude Test* (College Board Report. No. 81-3). New York, NY: College Entrance Examination Board.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Caulkins, J. P., Larkey, P. D., & Wei, J. (1996). *Adjusting GPA to reflect course difficulty*. Working paper, Heinz School of Public Policy and Management, Carnegie Mellon University.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.

Hoover, H. D., Roller, C., Liddell, D., Moore, J., McCarthy, C., & Hlebowitsh, P. (1999). *Grading practices in the college of education: A report of the committee on grading practices.*

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science, 12*(4), 251-278.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer.

Larzelere, R. E., & Mulaik, S. A. (1977). Single-sample tests for many correlations. *Psychological Bulletin, 84*(3), 557-569.

Linn, R. L. (1966). Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement, 3*(4), 313-329.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., & Jirele, T. J. (1994). Adjusting college grade point average criteria for variations in grading standards: A comparison of methods. *Journal of Applied Psychology, 79*(2), 178-183.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Thissen, D. (1991). *Multilog User's Guide – Version 6*. Chicago, IL: Scientific Software.

Thissen, D. (2000). *Multilog FAQs*. Available: http://www.ssicentral.com/irt/thissen.htm. (Retrieved on 7/3/00).

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

Wright, B. D., & Linacre, J. M. (1990). *A User's Guide to BIGSTEPS: Rasch-Model Computer Program Version 2.0*. Chicago, IL: MESA Press.

Young, J. W. (1990a). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement, 27*(2), 175-186.

Young, J. W. (1990b). Are validity coefficients understated due to correctable defects in the GPA? *Research in Higher Education, 31*(4), 319-325.

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement, 28*, 37-47.

Young, J. W. (1993). Grade Adjustment Methods. *Review of Educational Research, 63*(2), 151-165.

Ziomek, R. J. & Svec, J. C. (1995). *High school grades and achievement: evidence of grade inflation*. (ACT Research Report. No. 95-3). Iowa City, IA: American College Testing, Inc.

Appendix A

Formulas and Number of Course Parameters for Models Selected for Grade Adjustment

Abbreviation	Description	Formulas	No. of Course Parameters [a]
GPA	Grade point average	$x_{ij} = \theta_i + \varepsilon_{ij}$	0
Linear Models			
ADD	Additive	$x_{ij} = \theta_i - \delta_j + \varepsilon_{ij}$	$m-1$
MULT	Multiplicative	$x_{ij} = \left(\frac{1}{\alpha_j}\right)\theta_i + \varepsilon_{ij}$	$m-1$
COMB	Combined	$x_{ij} = \left(\frac{1}{\alpha_j}\right)\theta_i - \delta_j + \varepsilon_{ij}$	$2(m-1)$
Polytomous Item Response Models [b]			
RS	Rating Scale	$\pi_{nik} = \frac{\exp\left(\sum_{j=0}^{k}(\theta_n - \delta_i - \tau_j)\right)}{\sum_{x=0}^{m_i}\exp\left(\sum_{j=0}^{x}(\theta_n - \delta_i - \tau_j)\right)}$	$m+(k-1)$
PC	Partial Credit	$\pi_{nik} = \frac{\exp\left(\sum_{j=0}^{k}(\theta_n - \delta_{ij})\right)}{\sum_{x=0}^{m_i}\exp\left(\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right)}$	$m(k-1)$
GR	Graded Response	$\pi_{nik} = \frac{1}{1+\exp(\alpha_i(\theta_n - \tau_{ik-1}))} - \frac{1}{1+\exp(\alpha_i(\theta_n - \tau_{ik}))}$	$m+m(k-1)$
GPC	Generalized Partial Credit	$\pi_{nik} = \frac{\exp\left(\sum_{j=0}^{k}(\alpha_i(\theta_n - \delta_{ij}))\right)}{\sum_{x=0}^{m_i}\exp\left(\sum_{j=0}^{x}(\alpha_i(\theta_n - \delta_{ij}))\right)}$	$m+m(k-1)$

Note.

[a] Assume that there are k distinct grades within each course, and that there are m courses.

[b] For polytomous IRT models, π_{nik} represents the probability of person n getting grade k for course i; m_i represents the highest grade assigned for course i, where the lowest course grade is 0.

Figures

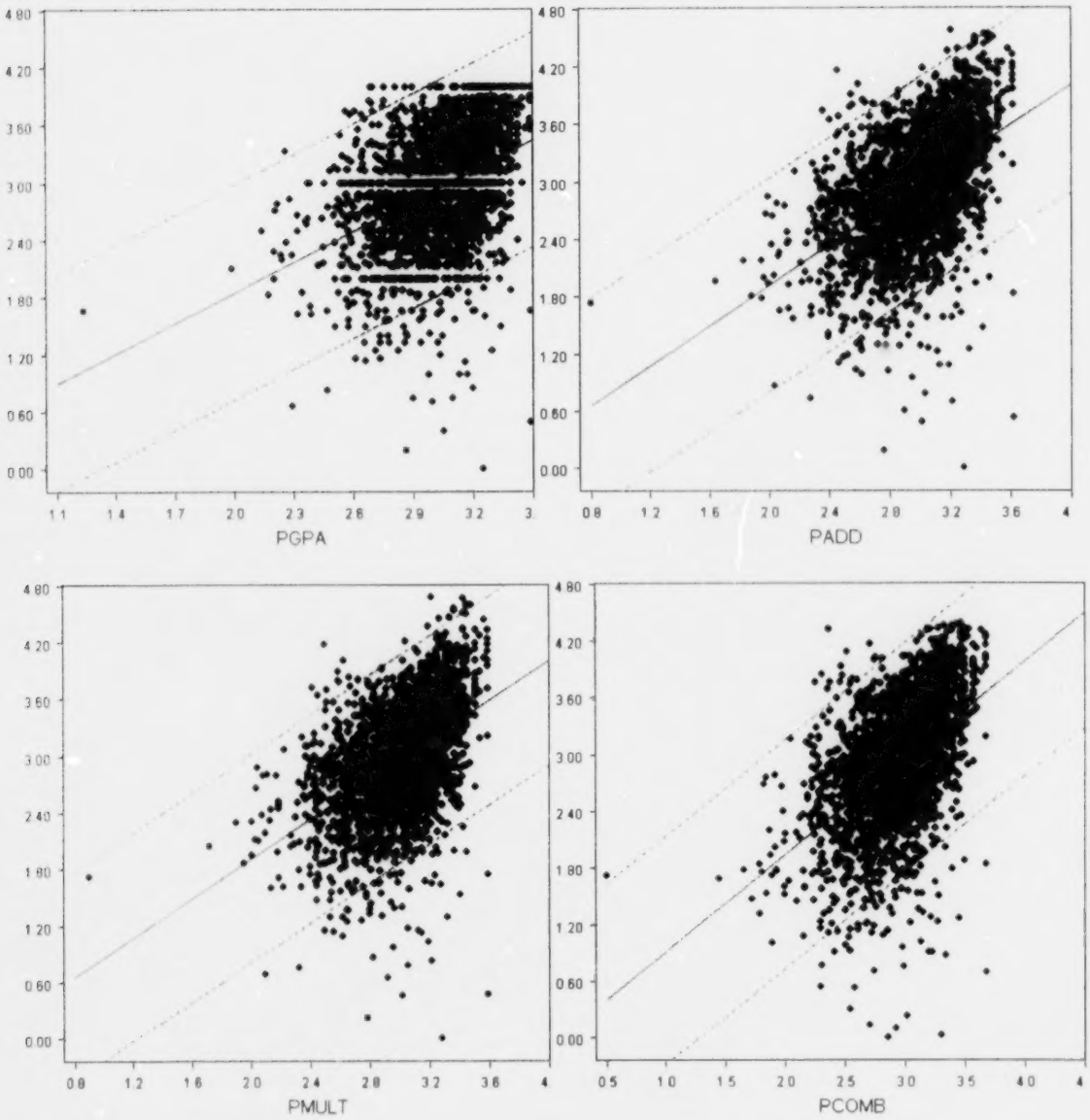

FIGURE 1. Scattered plot of achievement measures on predicted achievement measures from preadmission variables for: upper left – Unadjusted GPA, upper right – Additive adjusted GPA, lower left – Multiplicative adjusted GPA, lower right – Combined adjusted GPA.

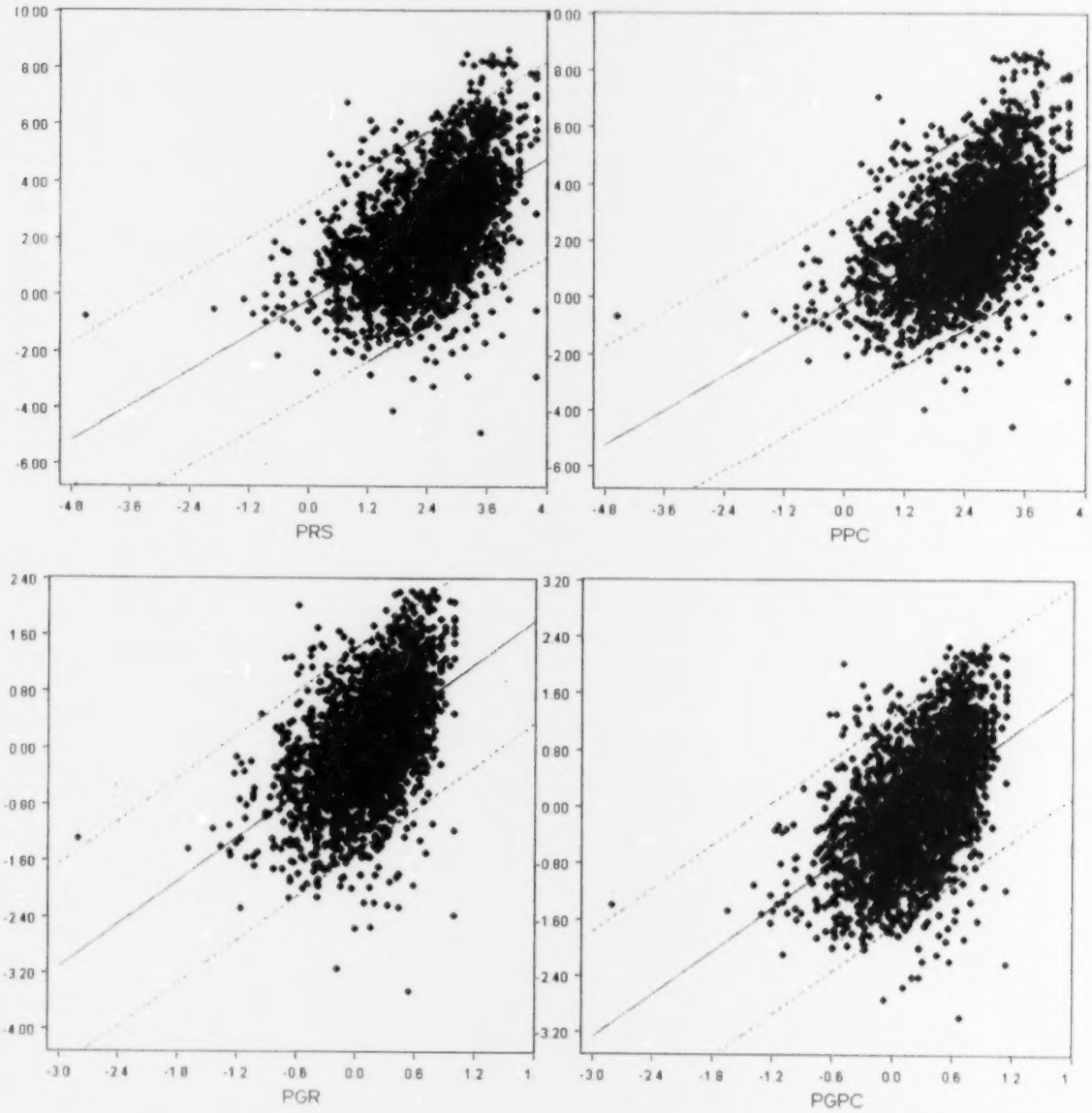

FIGURE 2. Scattered plot of achievement measures on predicted achievement measures from preadmission variables for: upper left – Rating Scale adjusted GPA, upper right – Partial Credit adjusted GPA, lower left – Graded Response adjusted GPA, lower right – Generalized Partial Credit adjusted GPA.

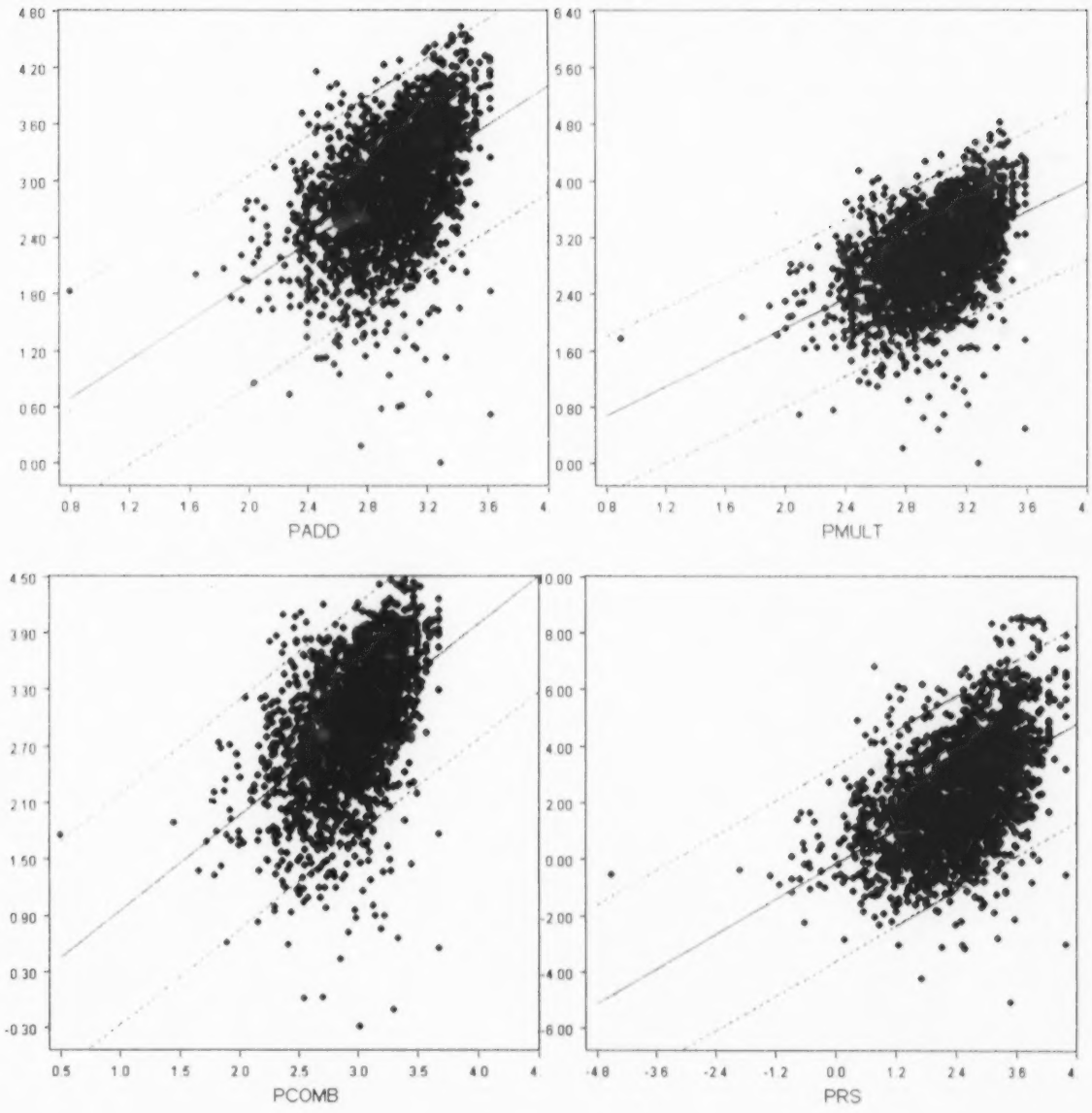

FIGURE 3. Scattered plot of cross-validated achievement measures on predicted achievement measures from preadmission variables for: upper left – Additive adjusted GPA, upper right – Multiplicative adjusted GPA, lower left – Combined adjusted GPA, lower right – Rating Scale adjusted GPA.

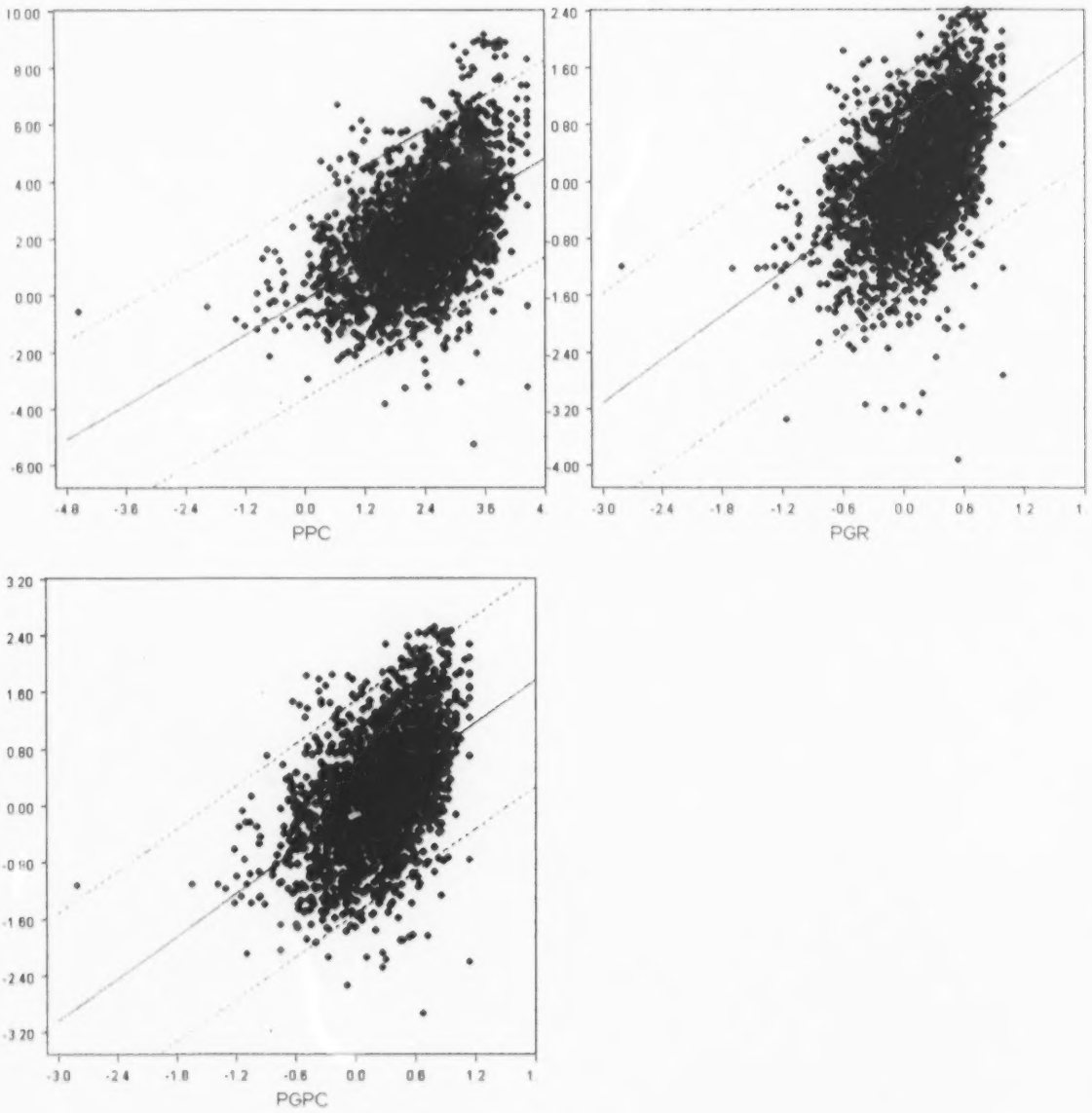

FIGURE 4. Scattered plot of crossed-validated achievement measures on predicted achievement measures from preadmission variables for: upper left – Partial Credit adjusted GPA, upper right – Graded Response adjusted GPA, lower left – Generalized Partial Credit adjusted GPA.